The Battle of the Neighborhoods - Report

IBM Data Science Professional Certificate - Capstone Project
Opening a new Restaurant in Brooklyn, USA
Francesco Morena - August 2019

Part 1 - A description of the problem and a discussion of the background.

Problem Background:

The City of New York, is the most populous city in the United States. New York City's demographics show that it is a large and ethnically diverse metropolis with a long history of international immigration.It is diverse and is the financial capital of USA. It is multicultural and with it's diverse culture, comes diverse food items.

There are many resturants in New york City, each beloning to different categories like Italian, Mexican, Puerto Rican, Indian etc.

So as part of this project, we will list, visualize and explor/analyze all major parts of New York City that has italian resturants.

Problem Description:

A restaurant is a business which prepares and serves food and drink to customers in return for money. The City of New York is famous for its excelllent cuisine. It's food culture includes an array of international cuisines influenced by the city's immigrant history.
So it is evident that to survive in such competitive market it is very important to startegically plan. Various factors need to be studied inorder to decide on the Location such as:

1. Are there any Farmers Markets, Wholesale markets etc nearby so that the ingredients can be purchased fresh to maintain quality and cost?
2. Are there any venues like Gyms, Entertainmnet zones, Parks etc nearby where floating population is high?
3. Who are the competitors in that location?
   etc...

The XYZ Company Ltd. need to choose the correct location to start its first venture. If this is successful they can replicate the same in other locations. First move is very important, thereby choice of location is very important.

My client, a XYZ Company successful restaurant chain in Italy is looking to expand operation into USA through New York City.

Since the NYC demography is so big, my client needs deeper insight from available data in other to decide where to establish the first America “palace” restaurant.
My client's preference is the Brooklyn area, where a good group of Italians live.

Target Audience:

The objective is to locate and recommend to the management which neighborhood of New York City will be best choice to start an Italian Restaurant.

Considering the diversity of the great metropolis, there is a high multicultural sense. This would interest anyone who wants to start a new activity in a great city as NY.

Success Criteria:

Which location should be suggested to the stakeholder? The success criteria of the project will be a good recommendation of borough/Neighborhood choice to the Company.

2. Data Requirements and Extraction:

One borough will be analysed in this project : ***Brooklyn***.

To solve this problem, we will need the following data:

- New York City data that contains list Boroughs, Neighborhoods along with their latitude and longitude.
  - Data source: https://cocl.us/new_york_dataset
  - Description: This data set contains the required information. And we will use this data set to explore various neighborhoods of new york city.
- Italian restaurants in each neighborhood of new york city.
  - Data source: Fousquare API
  - Description: By using this api we will get all the venues in each neighborhood. We can filter these venues to get only italian restaurants.

Data 1 : Newyork city geographical coordinates data will be utilized as input for the Four square API,that will be leveraged to provision venues information for each neighborhood. We will use the Foursquare API to explore neighborhoods in New York City. The below is image of the Foursquare API data.

```
# get new york data and create a DataFrame
new_york_data=get_new_york_data()
df = pd.DataFrame(new_york_data)
df.head()
```

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

Neighborhood has a total of 5 boroughs and 306 neighborhoods. In order to segement the neighborhoods and explore them, we will essentially need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the the latitude and logitude coordinates of each neighborhood.

3.Methodology :

Business Understanding :
Our main goal is to get optimum location for new restaurant business in Brooklyn, New York City, for my client.

Analytic Approach :
In general, this project would be encompassing a series of Data Science techniques, including, but not limited to, Data Cleaning, Data Wrangling and Machine Learning (K-Means clustering algorithm)
• Approach:
- o Collect the New York City data from https://cocl.us/new_york_dataset
- o Using FourSquare API we will find all venues for each neighborhood
- o Filter out all neighborhoods of Brooklyn
- o Filter out all venues that are Italian Restaurants, Office, University, Hischool
- o Analyze venues of interest using FouSquare API
- o Using rating for each neighborhoods , we will sort that data
- o Visualize the Ranking of neighborhoods using folium library(python)
- o K-Means clustering and Clustering Analisys

For each neighborhood, the sums of the office, school, university and italian restaurant were computed and for each of this 4 categories, a weight (or penalty) has been defined according to what stockholder considers the most important.

- o Italian restaurant have been weighted with -1, since stockholder wants to avoid concurrence
- o HighSchools have been weighted with 1, since employees/student are good customers
- o Universities have been weighted with 1.5, since employees/students are good customers
- o Offices have been weighted with 2, since employees are even better customers
- o A score was computed for each locality as the weighted sum of the number of venues in each of the 4 categories (school, university, office, Italian Restaurant)
- o Lastly, K-Means clustering is performed on this data set to return clusters, or categories of neighborhoods in terms of number of Italian Restaurant

Questions that can be asked using the aforementioned datasets:
• What is best location in Brooklyn that has potential Italian Restaurant market?
• Which areas are saturated with Italian restaurants, which untapped?

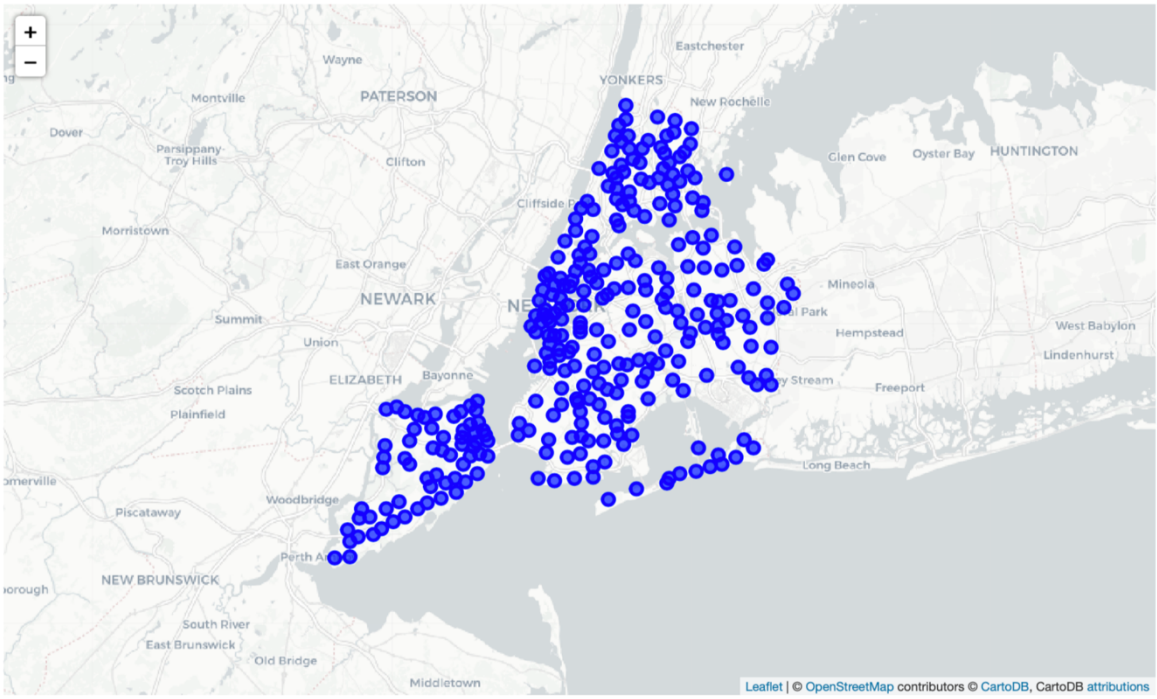
Exploratory Data Analysis :

Data 1- New york city Geographical Coordinates Data.

1. In this we load the data and explore data from New York file.
2. Transform the data of nested python dictionaries into a pandas dataframe.
3. This dataframe contains the geographical coordinates of New York city neighborhoods.

4. This data will used to get Venues data from FourSquare.
5. We used geopy and folium libraries to create a map of New York City with neighborhoods superimposed on top.

New York neighborhoods visualization

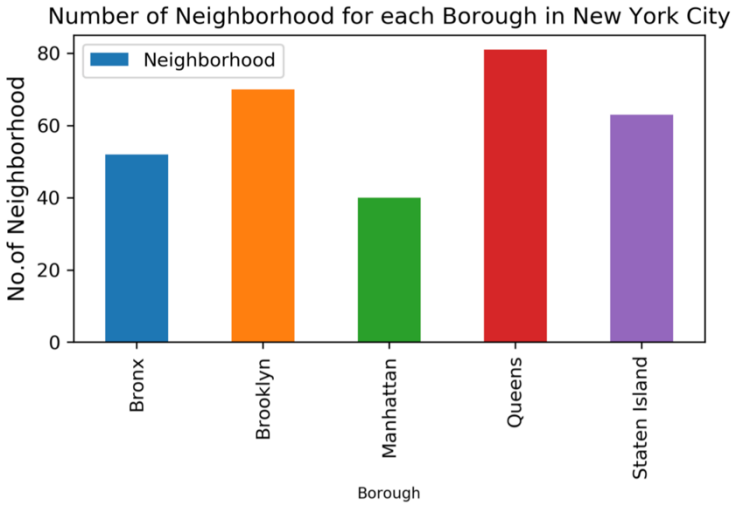


There is a total of 5 boroughs and 306 neighborhoods in New York City

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

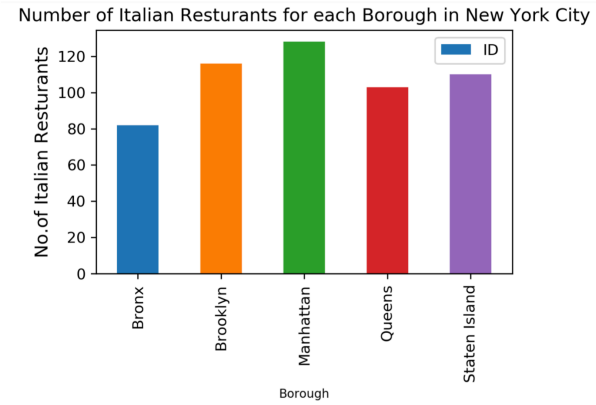
```
df.shape
```

```
(306, 4)
```



Queens has the highest number of neighborhoods (80), followed by Brooklyn (70) and Staten Island (60).

Data 2- Second data which is used is the FourSquare created dataset with italian restaurant. In this we will be using the data of Italian restaurant data. There are totally 539 Italian restaurants in New York city. Highest number are in Manhattan and Brooklyn. And lowest in Staten Island, Queens, and Bronx. The proof of this is as given below.




```
italian_rest_ny.head()
```

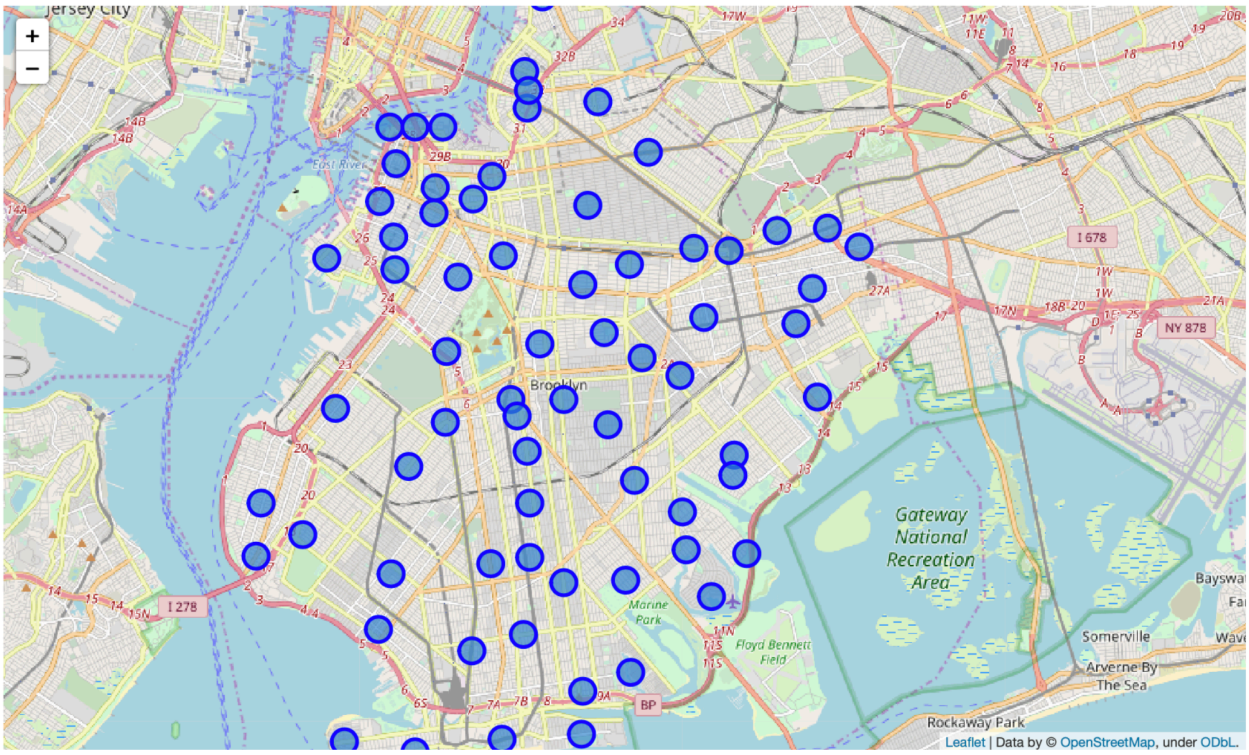
	Borough	Neighborhood	ID	Name
0	Bronx	Riverdale	55aaee4d498e3cbb70e625d6	Bella Notte Pizzeria
1	Bronx	Kingsbridge	55aaee4d498e3cbb70e625d6	Bella Notte Pizzeria
2	Bronx	Woodlawn	511edb6de4b0d58346fd272d	Patrizia's Of Woodlawn
3	Bronx	Woodlawn	4d3cb3026b3d236a066a6364	Rivers Edge
4	Bronx	Baychester	4c9518076b35a143d5dc21dc	Fratelli's

```
italian_rest_ny.shape
```

```
(539, 4)
```

Due to the aim of the project and the request of the stockholder, we taking into account Italian Restaurants in a specific area: Brooklyn.

Brooklyn neighborhoods visualization



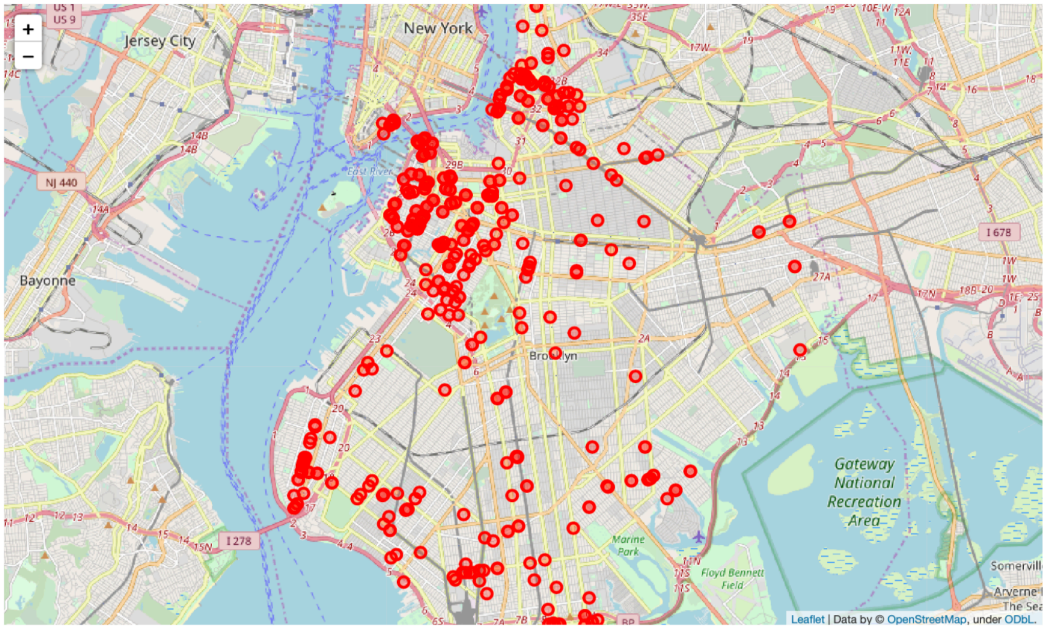
From Foursquare we will need following venues data:

- the Italian Restaurant venues of the Brooklyn
- the offices venues of Brooklyn
- the high schools venues of Brooklyn
- the universities venues of Brooklyn

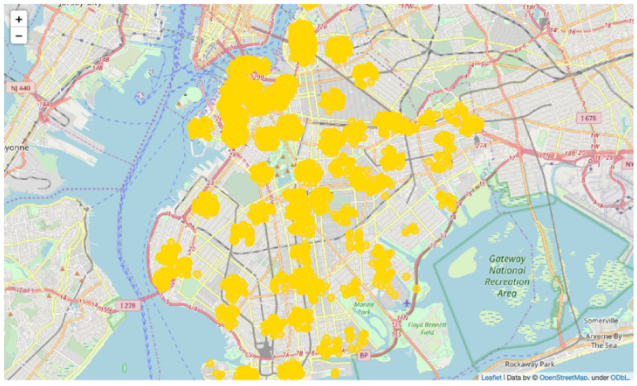
Using the geographical coordinates of each neighborhood foursquare API calls are made to get top 100 venues in a radius of 500 meters.

We used geopy and folium libraries to create a map to visualise venues of interest in Brooklyn, New York city.

Map of Italian Restaurants in Brooklyn



Map of Offices in Brooklyn

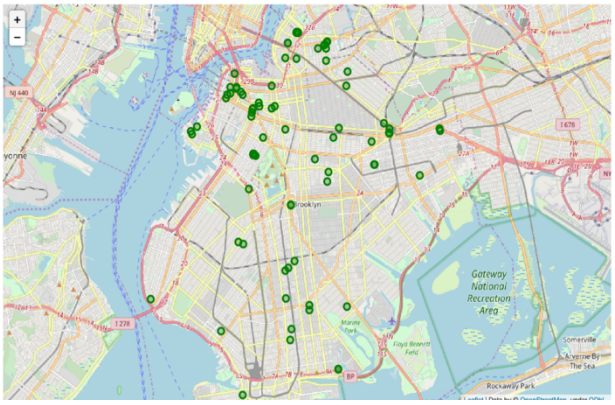


	Neighborhood	NeighborhoodLatitude	NeighborhoodLongitude	VenueName
0	Bay Ridge	40.625801	-74.030621	WishWash Carpet Cleaning & Restoration of Bay ...
1	Bay Ridge	40.625801	-74.030621	R&G Brenner Income Tax
2	Bay Ridge	40.625801	-74.030621	Integrated Financial Services
3	Bay Ridge	40.625801	-74.030621	Polizzi Fine Art
4	Bay Ridge	40.625801	-74.030621	H And R Block

```
broo_venues_office.shape
```

```
(3278, 7)
```

Map of high schools in Brooklyn




```
broo_venues_highschools.head()
```

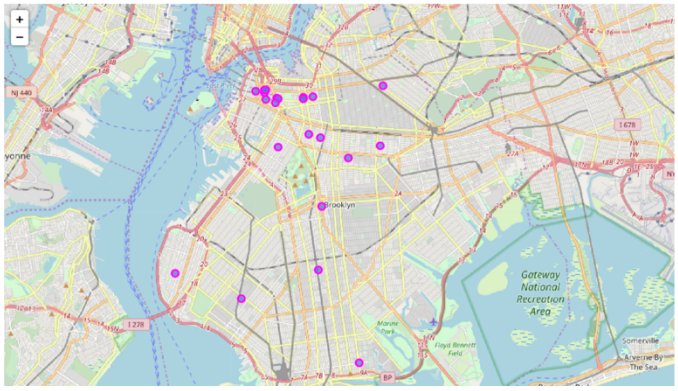
	Neighborhood	NeighborhoodLatitude	NeighborhoodLongitude	Venue
0	Manhattan Terrace	40.614433	-73.957438	Imagine Aca
1	Crown Heights	40.670829	-73.943291	St. Marks S
2	Windsor Terrace	40.656946	-73.980073	Bishop Ford Central Catholic High S
3	Prospect Heights	40.676822	-73.964859	Acorn Community High S
4	Brownsville	40.663950	-73.910235	Teachers Preparatory High S

```
broo_venues_highschools.shape
```

```
(81, 7)
```

There are 81 **high schools** across Brooklyn

Map of Universities in Brooklyn



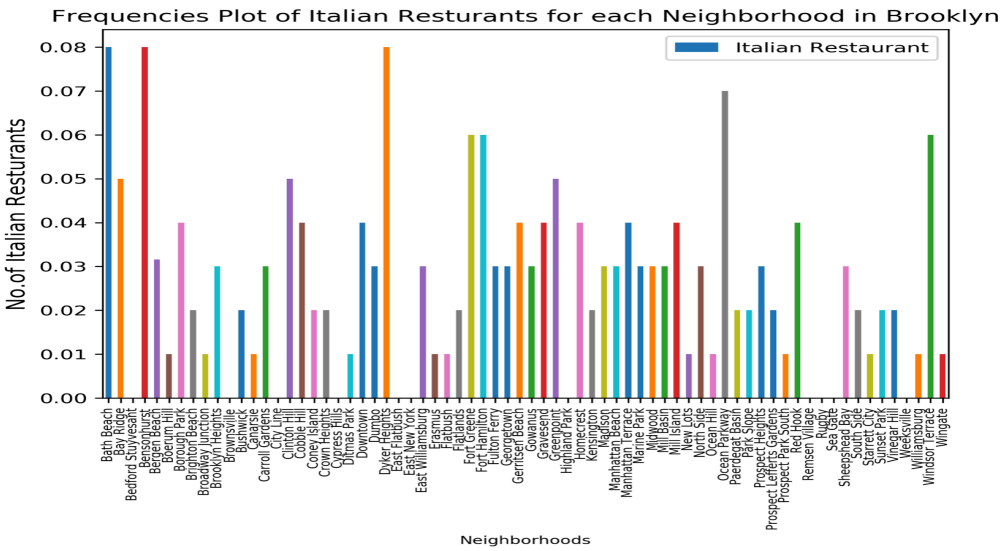
	Neighborhood	NeighborhoodLatitude	NeighborhoodLongitude	
0	Bay Ridge	40.625801	-74.030621	Stepping Sto
1	Bensonhurst	40.611009	-73.995180	In
2	Sheepshead Bay	40.586890	-73.943186	Financ
3	Crown Heights	40.670829	-73.943291	Cornell University Cooperative
4	Prospect Heights	40.676822	-73.964859	Bacca

```
broo_venues_uni.shape
```

```
(24, 7)
```

There are 24 **Universities** across Brooklyn

The next step was to calculate the frequencies of Italian Restaurants for each neighborhood in Brooklyn as given below.



RESULTS :

For each neighborhood, the sums of the office, school, university and italian restaurant were computed and for each of this 4 categories, a weight (or penalty) has been defined according to what stockholder considers the most important.

o Italian restaurant have been weighted with -1, since stockholder wants to avoid concurrence

o HighSchools have been weighted with 1, since employees/student are good customers

o Universities have been weighted with 1.5, since employees/students are good customers

o Offices have been weighted with 2, since employees are even better customers

o A score was computed for each locality as the weighted sum of the number of venues in each of the 4 categories (school, university, office, Italian Restaurant)

o Lastly, K-Means clustering is performed on this data set to return clusters, or categories of neighborhoods in terms of number of Italian Restaurant

	Neighborhood	Score
86	Downtown	195.5
283	Dumbo	193.0
49	Greenpoint	191.0
64	Brooklyn Heights	184.5
70	Park Slope	181.5
96	North Side	176.0
87	Boerum Hill	168.5
280	Vinegar Hill	165.0
99	Fort Hamilton	163.0
53	Manhattan Terrace	160.0
61	Williamsburg	154.0

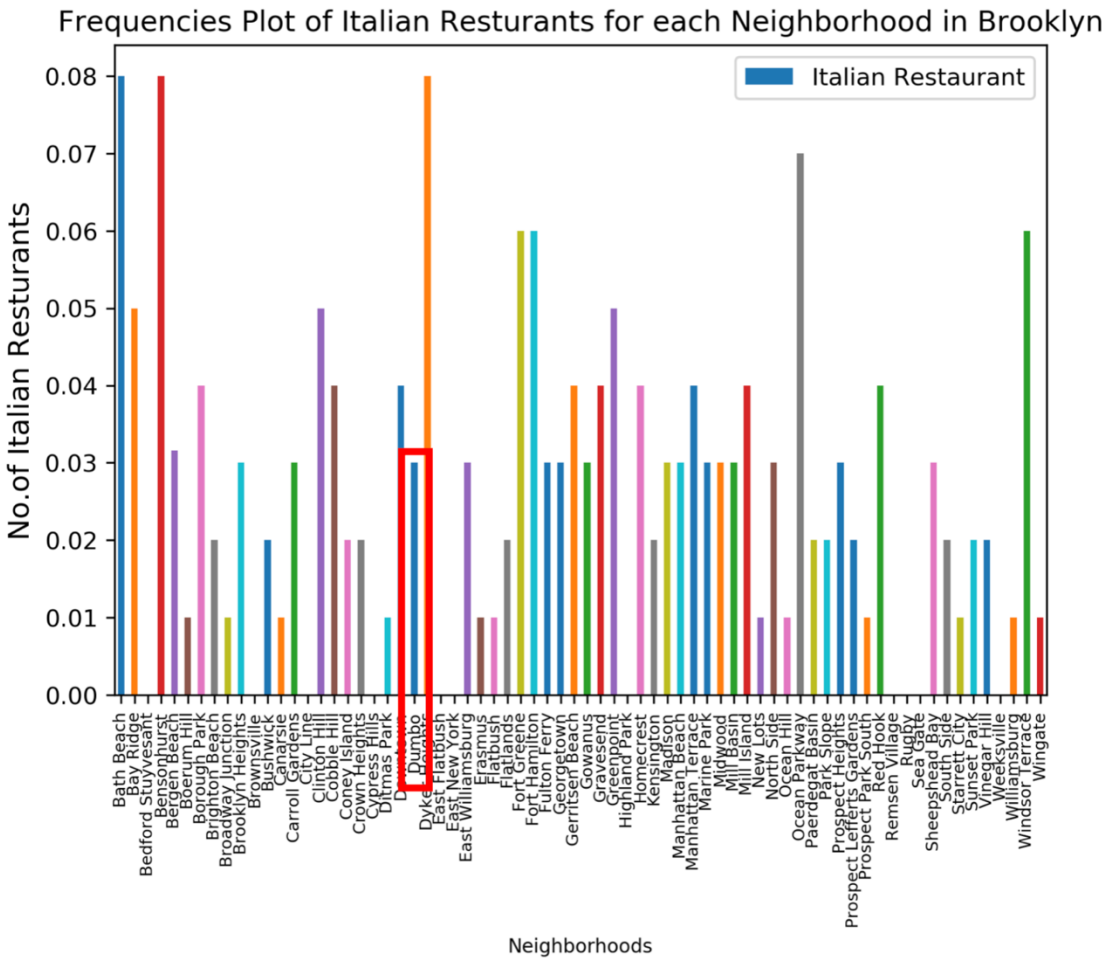
The neighborhoods with the best score are:

Downtown, Dumbo, Greenpoint, Brooklyn Heights and Park Slope are some of the best neighborhoods for stockholder in order to open his Italian Restaurant.

These options maximize the number of potential customers from Offices and Universities and at the same time have not too large competitor.

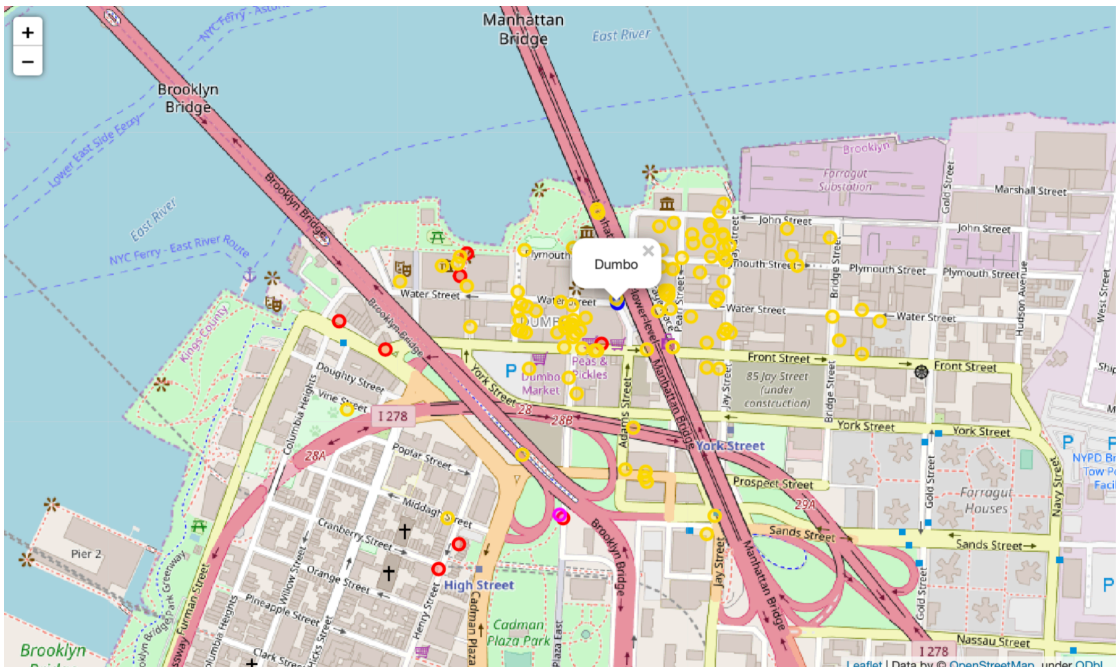
However, analyzing the Neighborhood with the best score, Dumbo would seem to be the most suitable as the one with the lowest number of Italian restaurants (7), compared to the Downtown (17), Greenpoint (9), Broklyn Heights (22) and Park Slope (26).

Dumbo with a score of 193.0 being the best option



Best Place for the italian restaurant in Broklyyn is ‘Dumbo’.

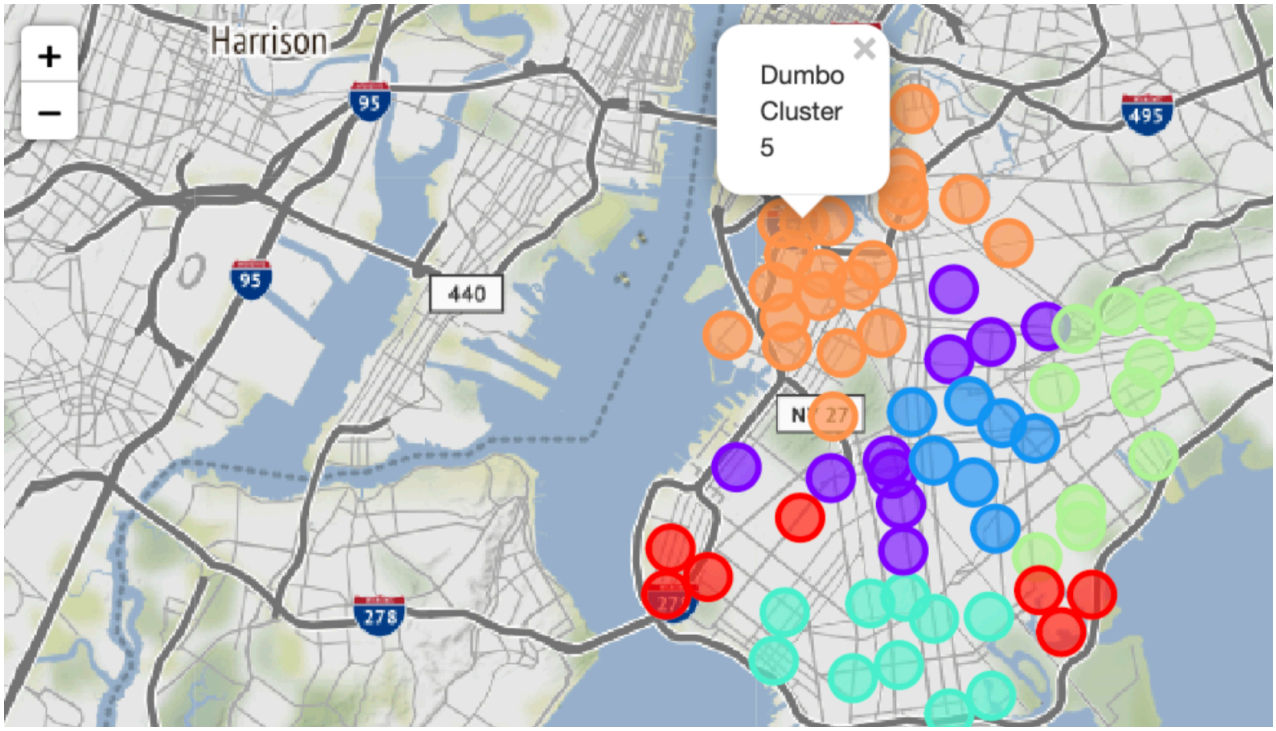
Red= Italian Restaurants Gold= Offices



K-Means Clustering:

To cluster the neighborhoods into six clusters we used the K-Means clustering Algorithm. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. It uses iterative refinement approach.

In the below Map Visualization, we can see the different types of clusters created by using K-Means for Brooklyn



Top 5 most common venues in cluster 0

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
count	10.0	NaN	NaN	NaN	NaN	NaN
freq	NaN	50.00%	20.00%	20.00%	20.00%	20.00%
mode	NaN	Pizza Place	Italian Restaurant	Beach	Bakery	Bar

Top 5 most common venues in cluster 1

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
count	10.0	NaN	NaN	NaN	NaN	NaN
freq	NaN	40.00%	30.00%	20.00%	30.00%	20.00%
mode	NaN	Caribbean Restaurant	Pizza Place	Café	Coffee Shop	Café

Top 5 most common venues in cluster 2

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
count	7.0	NaN	NaN	NaN	NaN	NaN
freq	NaN	100.00%	28.57%	28.57%	42.86%	28.57%
mode	NaN	Caribbean Restaurant	Café	Donut Shop	Pizza Place	Juice Bar

Top 5 most common venues in cluster 3

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
count	11.0	NaN	NaN	NaN	NaN	NaN
freq	NaN	54.55%	27.27%	18.18%	36.36%	18.18%
mode	NaN	Pizza Place	Bagel Shop	Pizza Place	Bagel Shop	Grocery Store

Top 5 most common venues in cluster 4

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
count	11.0	NaN	NaN	NaN	NaN	NaN
freq	NaN	72.73%	27.27%	36.36%	27.27%	27.27%
mode	NaN	Pizza Place	Discount Store	Supermarket	Fast Food Restaurant	Sandwich Place

Top 5 most common venues in cluster 5

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
count	21.0	NaN	NaN	NaN	NaN	NaN
freq	NaN	23.81%	28.57%	23.81%	19.05%	14.29%
mode	NaN	Bar	Pizza Place	Coffee Shop	Italian Restaurant	Cocktail Bar

The following are the highlights of the 6 clusters:

- Pizza Place and Italian Restaurant are the most popular in the cluster 0. The Italian Restaurants are very popular in this cluster, especially in Dyker Heights, Fort Hamilton and Bay Ridge areas;
- In clusters 1 and 2 predominant is the Caribbean Restaurant, which suggests a prevalent Caribbean presence, followed by Cafe and Pizza Place;
- Pizza Place, Bagel shop, and Grocery Store are the most popular in the cluster 3;
- Pizza Place, Discount Store, Supermarket, and Fast Food Restaurant are the most popular in the cluster 4;
- Bar, Pizza Place, Coffee Shops are the most popular in the cluster 5.

Conclusion and future directions

Cluster 5 is the most viable clusters to create a brand Italian Restaurant. Their proximity to other amenities is paramount. In this cluster, even if there are other Italian restaurants, the market is not saturated. So, the rivalry could be minimal.

In conclusion, this project would have had better results if there were more data in terms of crime data within the area, parking access, bus and metro stop, and allowance of more venues exploration with the Foursquare (limited venues for free calls). Also, getting the ratings and feedbacks of the current restaurants within the clusters would have helped in providing more insight into the best location.